RT-06S Speaker Diarization Results and Speech Activity Detection Results

http://www.nist.gov/speech/tests/rt/rt2006/spring/

Jonathan Fiscus, John Garofolo, Jerome Ajot, Martial Michel

May 3, 2006

Rich Transcription 2006
Spring Meeting Recognition Workshop
at MLMI 2006

Outline

- Reference generation
- Diarization “Who Spoke When” results
  - Experiments with forced-alignment mediated references
- Diarization “Speech Activity Detection” results
- Proposals for next year

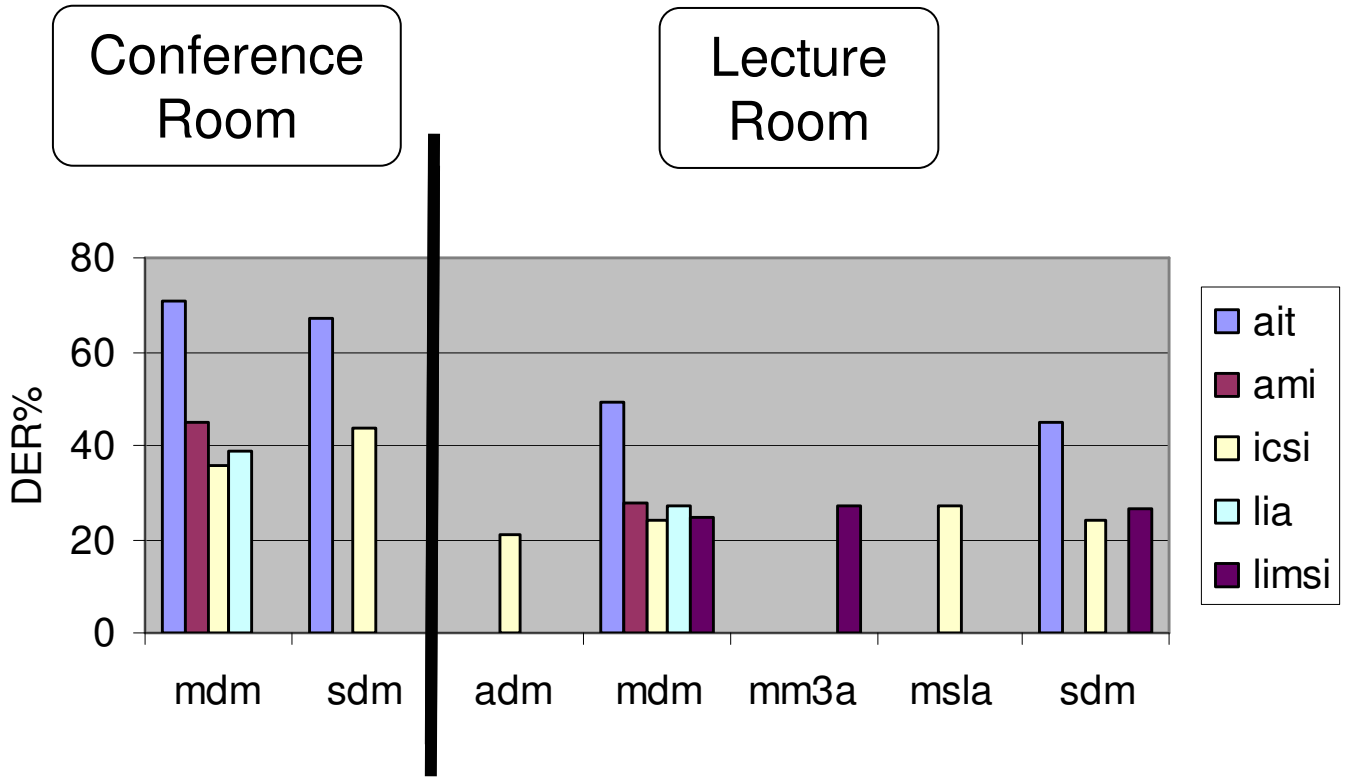
Diarization "Who Spoke When" (SPKR) Task

- Task definition
  - Identify the number of participants in each meeting and create a list of speech time intervals for each such participant
- Several input conditions:
  - Conference Room:
    - **MDM(primary)**, SDM, ADM, IHM
  - Lecture Room:
    - **MDM(primary)**, MM3A, MSLA, ADM, SDM
- Five participating sites:
  - AIT, AMI, ICSI/SRI, LIA, LIMSI

SPKR System Evaluation Method

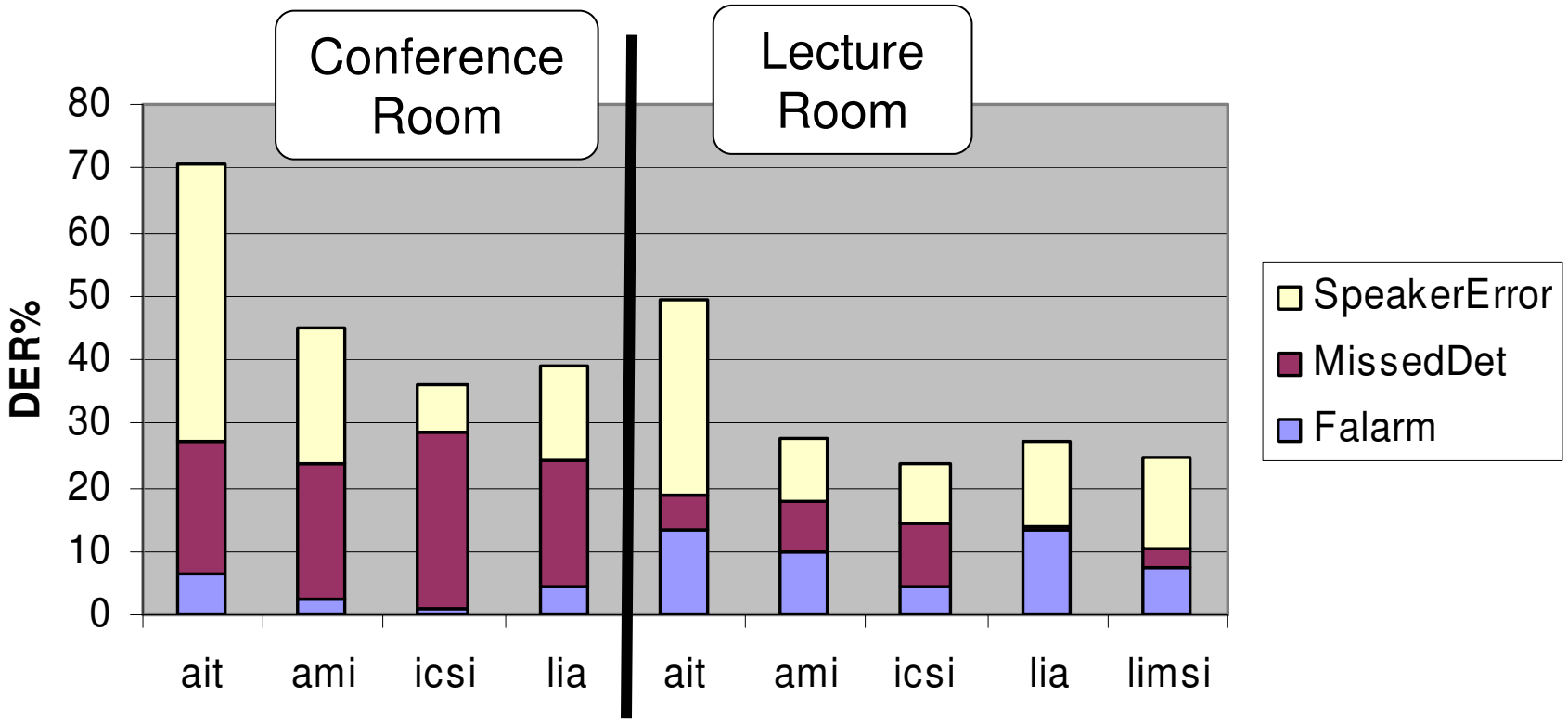
- Primary Metric
  - **Diarization Error Rate (DER) –** the ratio of incorrectly detected speaker time to total speaker time
    - System output speaker segment sets are mapped to reference speaker segment sets so as to minimize the total error
    - Errors consist of:
      - Speaker assignment errors (i.e., detected speech but not assigned to the right speaker)
      - False alarm detections
      - Missed detections
- Systems were scored using the mdeval tool
  - Forgiveness collar of +/- 250ms around reference segment boundaries
- DER on overlapping speech is the primary metric
  - Last year it was DER for non-overlapping speech
- Reference generation different than last year
  - Non-lexemes (speaker generated non-words e.g., laugh, cough) were stripped from the reference prior to reference file generation

RT-06S SPKR Results
Primary Systems, All Speech



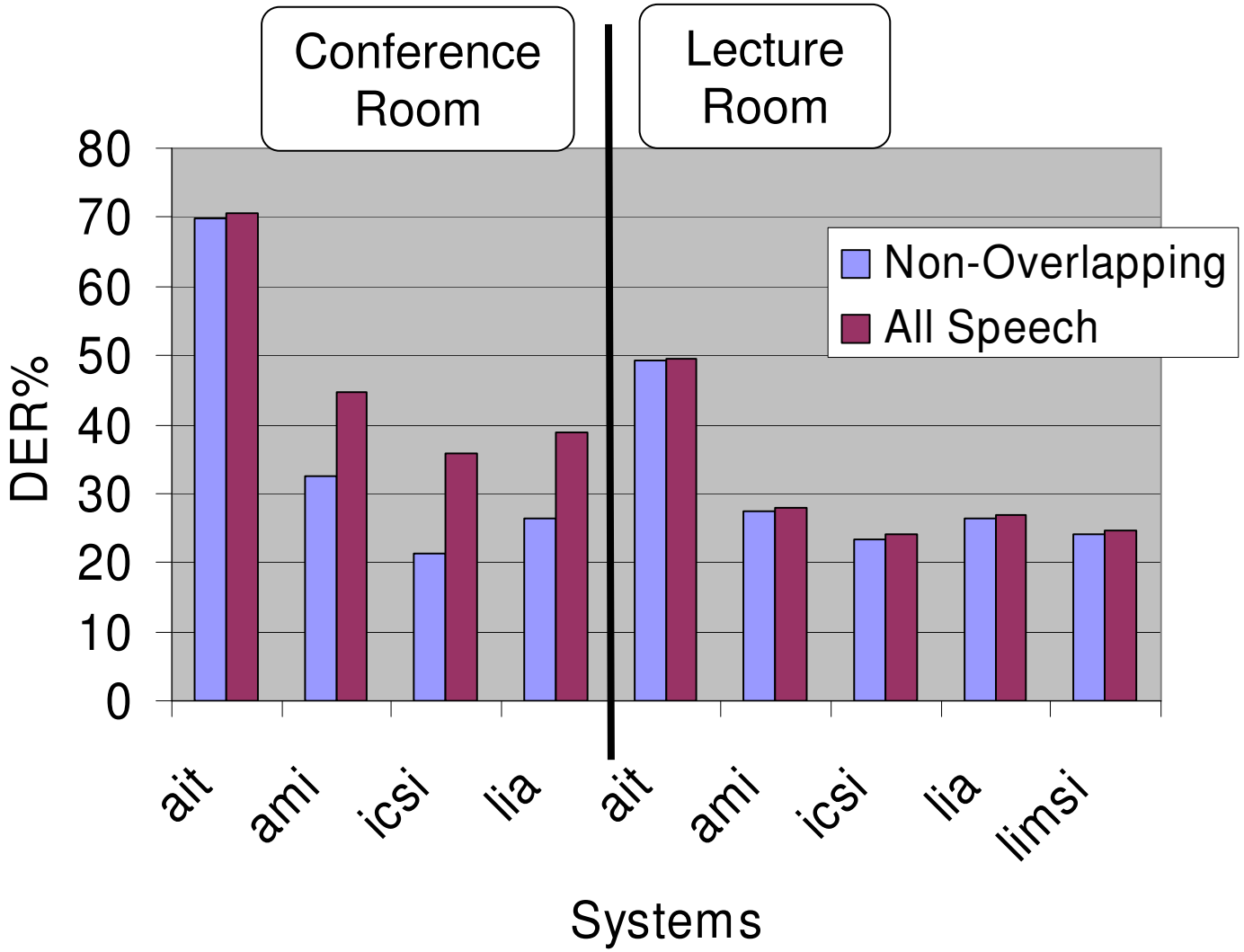
- Like 2005, Lecture data has lower error rates than conference data
- ICSI's ADM result is lower than their MDM result

RT-06S Primary SPKR MDM Systems
DER Split by Error Type



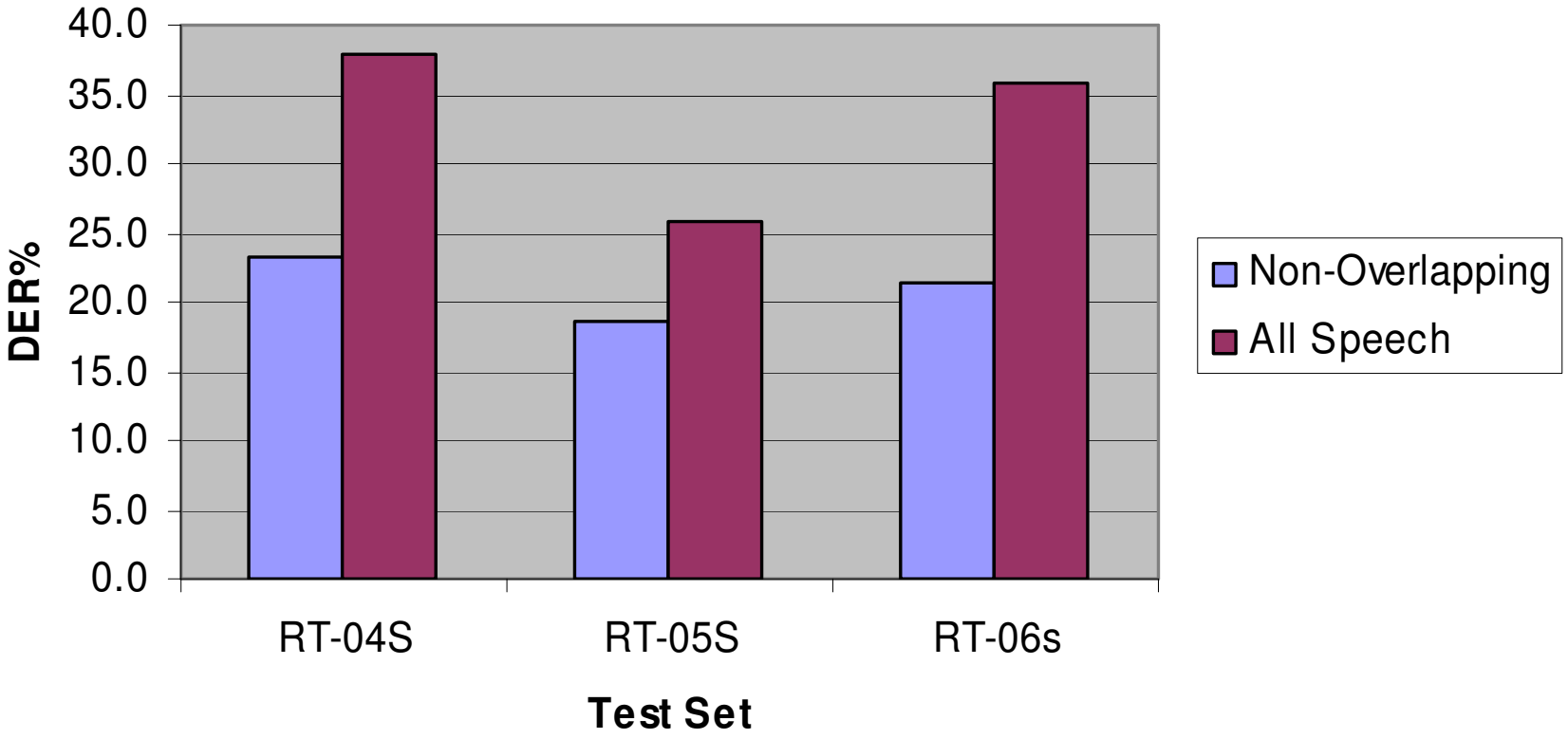
- Speaker Error and Missed Detections account for most of the error

RT-06S Primary MDM Systems
Non-overlapping Speech vs. All Speech



NIST National Institute of Standards and Technology

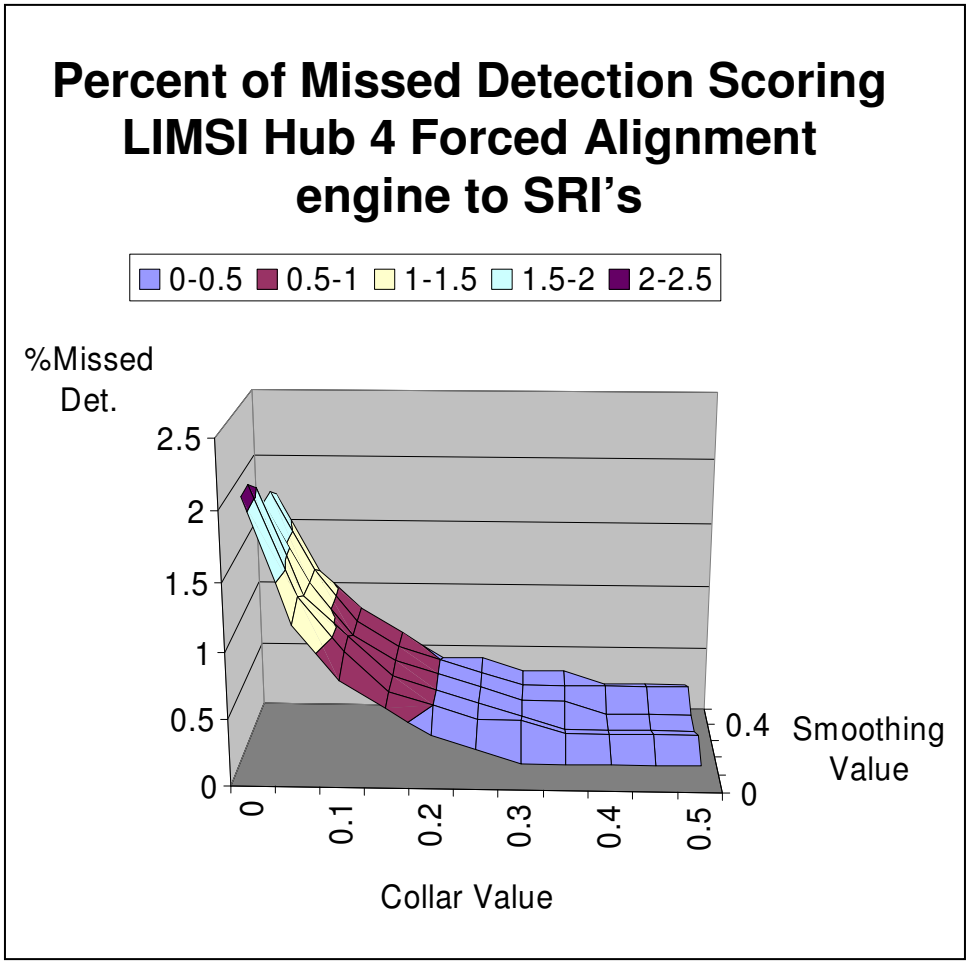
Historical Best System MDM SPKR Performance on Conference Data



- Lowest error rates are higher than last year
- Changes in reference may be the cause
  - We need a better mechanism for reference generation

Forced Word Alignment Mediated SPKR References

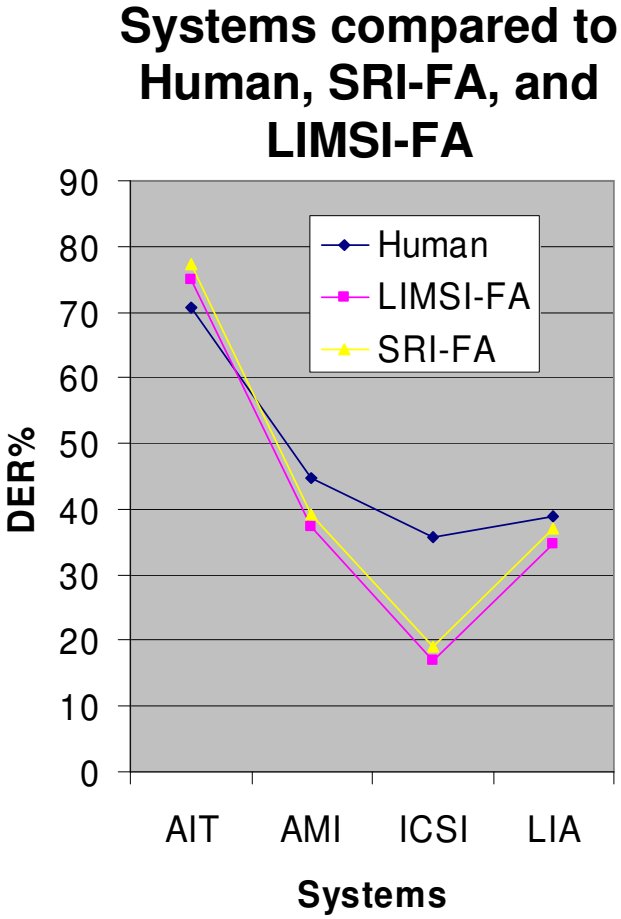
- EARS Program used force word alignments to generate reference segmentations for SPKR evaluation
- Advantages:
  - References will have consistent bias
- Disadvantages:
  - Forced aligners typically don't handle non-lexemes
  - Whose aligner to use?
  - How will it change the task?
- We studied two sets of forced alignments
  - SRI and LIMSI Hub 4(circa 2004)
  - Can an appropriate collar be determined?

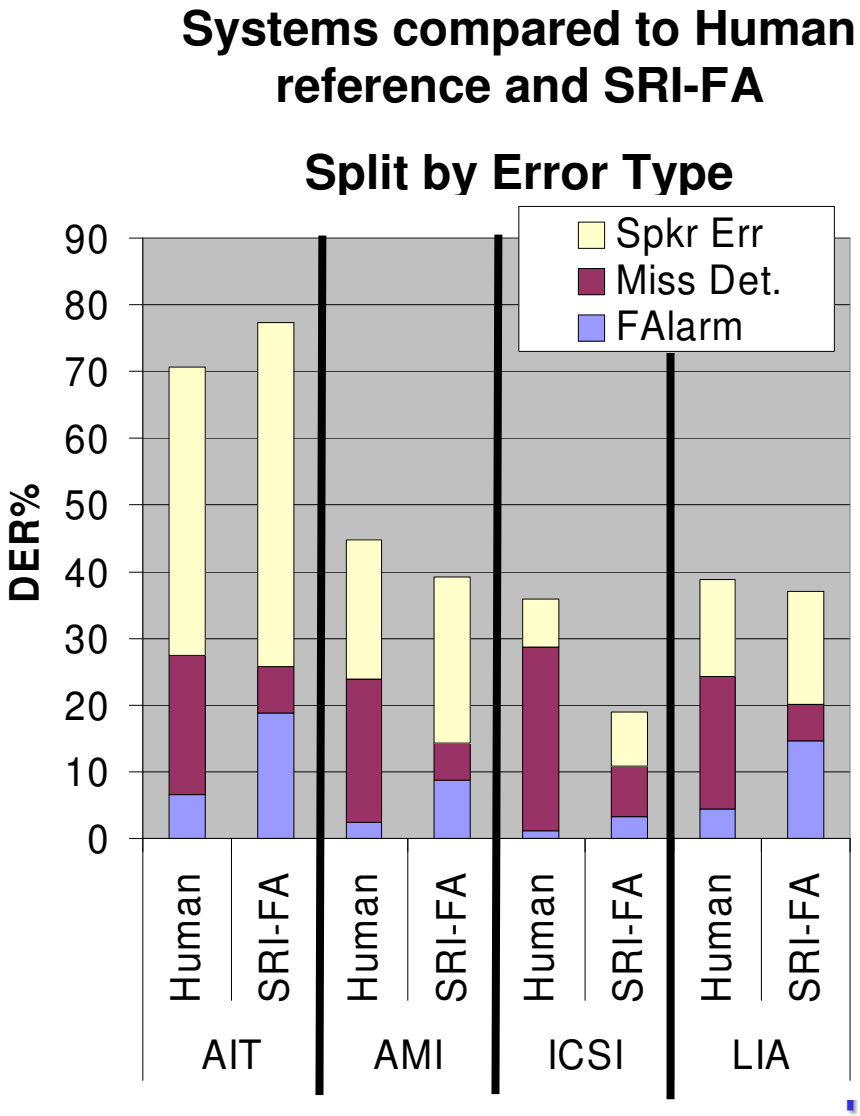


Collar of 0.25 homogenizes the data

NIST National Institute of Standards and Technology

RT-06S Primary MDM SPKR Systems
Alternative Reference Files





NIST National Institute of Standards and Technology

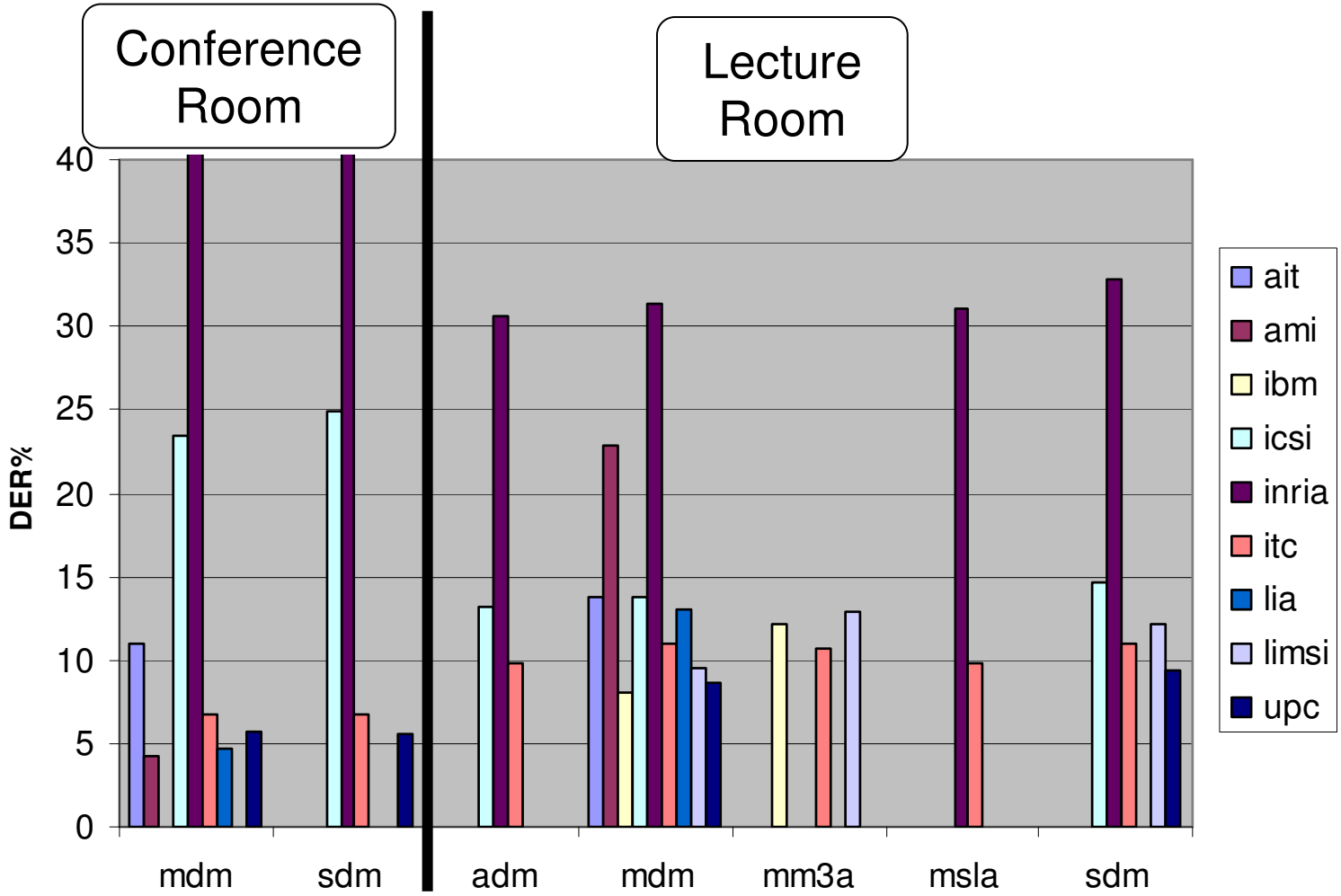
Diarization "Speech Activity Detection" (SAD) Task

- Task definition
  - create a list of speech time intervals where at least one person is talking
- Several input conditions:
  - Conference Room:
    - **MDM(primary)**, SDM, ADM, IHM
  - Lecture Room:
    - **MDM(primary)**, MM3A, MSLA, ADM, SDM
- Nine participating sites:
  - AIT, AMI, ICSI, IBM, INRIA, ITC, LIA, LIMSI, UPC

SAD System Evaluation Method

- Primary metric
  - Diarization Error Rate (DER)
    - Same formula and software as used for the SPKR task
    - Reduced to a two-class problem: speech vs. non-speech
    - No speaker assignment errors, just false alarms and missed detections
  - Forgiveness collar of +/- 250ms around reference segment boundaries

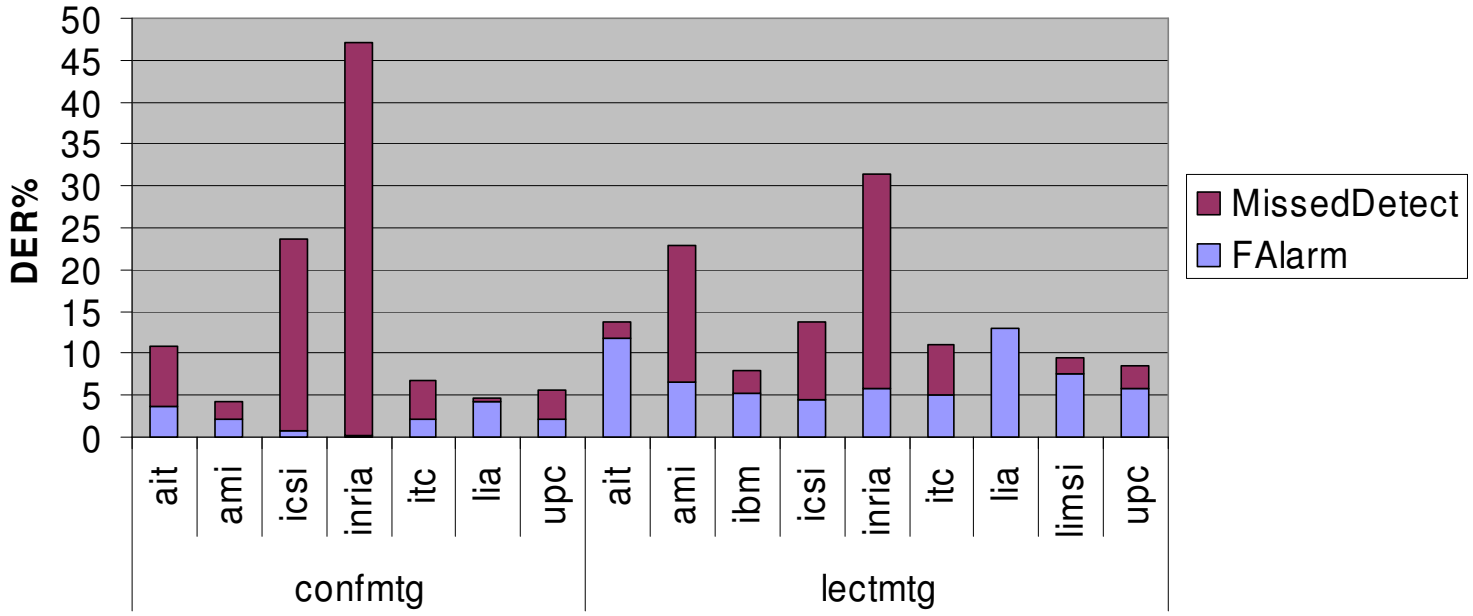
RT-06S SAD Results
Primary Systems



- Lecture data systems have higher error rates

NIST National Institute of Standards and Technology

RT-06S SAD Primary MDM Results Split by Error Type



- Low error rate systems have good balance in error types

Proposed SPKR/SAD Changes for RT-07

- Study the impact of changing to forced alignment files
  - How will this impact the task?
  - Are the forced alignments better or giving lower error rates?
  - Do the same references work for SAD?
- Use force alignment generated reference files
  - Re-score RT-05 with forced alignments